REPORT DOCUMENTATION

AD-A260 715

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATE COVERED
	September 1992	Journal Article

4. TITLE AND SUBTITLE
Assessing Semantic Knowledge Using Computer-based and Paper-based Media

5. FUNDING NUMBERS
None

6. AUTHOR(S)
P-A. Federico

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Navy Personnel Research and Development Center
San Diego, California 92152-6800

8. PERFORMING ORGANIZATION REPORT NUMBER
JN-92-09

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING
Computers in Human Behavior:
Vol. 8, pp. 169-191

11. SUPPLEMENTARY NOTES

14 pp

93-01021

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT *(Maximum 200 words)*

Using a within-subjects design, 75 naval pilots and flight officers were administered computer-based and paper-based tests to assess semantic knowledge in order to determine the relative reliabilities and validities of these two measurement modes. Estimates of internal consistencies, equivalences, and discriminative validities were computed for multiple performance measures. It was revealed that the relative reliabilities derived for these two assessment schemes using multivariate measurement criteria were not significantly different, and the discriminant validity of computer-based measures was superior to paper-based measures.

14. SUBJECT TERMS
Computer-based testing, paper-based testing, multivariate measurement criteria, discriminant validity

15. NUMBER OF PAGES
13

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

Computers in Human Behavior, Vol. 8, pp. 169–181, 1992
Printed in the U.S.A.

0747-5632/92 $5.00 + .00

Assessing Semantic Knowledge Using Computer-Based and Paper-Based Media

Pat-Anthony Federico

Navy Personnel Research and Development Center

Abstract — *Using a within-subjects design, 75 naval pilots and flight officers were administered computer-based and paper-based tests to assess semantic knowledge in order to determine the relative reliabilities and validities of these two measurement modes. Estimates of internal consistencies, equivalences, and discriminative validities were computed for multiple performance measures. It was revealed that the relative reliabilities derived for these two assessment schemes using multivariate measurement criteria were not significantly different, and the discriminant validity of computer-based measures was superior to paper-based measures.*

The consequences of computer-based assessment on examinees' performance are not obvious. The investigations that have been conducted on this topic have produced mixed results. Some studies (D. F. Johnson & Mihal, 1973; Serwer & Stolurow, 1970) demonstrated that test-takers do better on verbal items given by computer than on paper-based items; however, just the opposite was found by other studies (D. F. Johnson & Mihal, 1973; Wildgrube, 1982). One investigation (Sachar & Fletcher, 1978) yielded no significant differences resulting from computer-based and paper-based modes of administration on verbal items. Two studies (English, Reckase, & Patience, 1977; Hoffman & Lundberg, 1976) demonstrated that these two testing modes did not affect performance on memory-retrieval items. Sometimes (D. F. Johnson & Mihal, 1973) test-takers do better on quantitative tests when computer given, sometimes (Lee, Moreno, & Sympson, 1984) they do worse, and other times (Wildgrube, 1982) it may make no difference. Other studies have supported the equivalence of computer-based and paper-based administration (Elwood & Griffin, 1972; Hedl, O'Neil, & Hansen, 1973; Kantor, 1988; Lukin, Dowd, Plake, & Kraft, 1985). Some researchers (Evan & Miller, 1969; Koson,

Opinions or assertions contained herein are those of the author, and are not to be construed as official or reflecting the views of the Department of the Navy.

Requests for reprints should be sent to the author at Code 13, Navy Personnel Research and Development Center (NPRDC), San Diego, CA 92152-6800.

Kitchen, Kochen, & Stodolosky, 1970; Lucas, Mullin, Luna, & McInroy, 1977; Lukin et al., 1985; Skinner & Allen, 1983) have reported comparable or superior psychometric properties of computer-based assessment relative to paper-based assessment in clinical settings.

Investigations of computer-based presentation of personality items have yielded reliability and validity indices comparable to typical paper-based presentation (Katz & Dalby, 1981; Lushene, O'Neil, & Dunn, 1974). No significant differences were found in the scores of measures of anxiety, depression, and psychological reactance due to computer-based and paper-based administration (Lukin et al., 1985). Studies of cognitive tests have provided inconsistent findings, with some (Hitti, Riffer, & Stuckles, 1971; Rock & Nolen, 1982) demonstrating that the computerized version is a viable alterative to the paper-based version. Other research (Hansen & O'Neil, 1970; Hedl et al., 1973; D. F. Johnson & White, 1980; J. H. Johnson & K. N. Johnson, 1981), though, indicated that interacting with a computer-based system to take an intelligence test could elicit a considerable amount of anxiety which could affect performance.

Regarding computerized adaptive testing (CAT), some empirical comparisons (McBride, 1980; Sympson, Weiss, & Ree, 1982) yielded essentially no change in validity due to mode of administration. However, test-item difficulty may not be indifferent to manner of presentation for CAT (Green, Bock, Humphreys, Linn, & Reckase, 1984). When going from paper-based to computer-based administration, this mode effect is thought to have three aspects: (a) an overall mean shift where all items may be easier or harder, (b) an item-mode interaction where a few items may be altered and others not, and (c) a change in the nature of the task itself caused by computer administration. A computer simulation study (Divgi, 1988) demonstrated that a CAT version of the Armed Services Vocational Aptitude Battery had a higher reliability than a paper-based version for these subtests: (a) general science, (b) arithmetic reasoning, (c) word knowledge, (d) paragraph comprehension, and (e) mathematics knowledge. The inconsistent results of mode, manner, or medium of testing may be due to differences in methodology, test content, population tested, or the design of the study (Lee et al., 1984).

With computer costs coming down and people's knowledge of these systems going up, it becomes more likely economically and technologically that many benefits can be gained from their use. A direct advantage of computer-based testing is that individuals can respond to items at their own pace, thus producing ideal power tests. Some indirect advantages of computer-based assessment are increased test security, less ambiguity about students' responses, minimal or no paperwork, immediate scoring, and automatic records keeping for item analysis (Green, 1983a, 1983b). Some of the strongest support for computer-based assessment is based upon the awareness of faster and more economical measurement (Elwood & Griffin, 1972; D. F. Johnson & White, 1980; Space, 1981). Cory (1977) reported some advantages of computerized over paper-based testing for predicting job performance.

Ward (1984) stated that computers can be employed to augment what is possible with paper-based measurement (e.g, to obtain more precise information regarding a student than is likely with more customary measurement methods) and to assess additional aspects of performance. He enumerated and discussed potential benefits that may be derived from employing computer-based systems to administer traditional tests. Some of these are as follows: (a) individualized assessment, (b) increased flexibility and efficiency for managing test information, (c) enhanced economic value and manipulation of measurement databases, and (d) improved

diagnostic testing. Millman (1984) agreed with Ward about computer-based measurement encouraging individualized assessment and designing software within the context of cognitive science. Also, limiting computer-based assessment is not so much hardware inadequacy, but incomplete comprehension of the processes intrinsic to testing (Federico, 1980).

Simplistic conceptual or associative knowledge can be represented as semantic networks (Barr & Feigenbaum, 1981). These symbolic schemes usually consist of nodes (e.g., circles or boxes) and links (e.g., arcs or arrows) connecting the nodes. Typically, nodes represent objects, concepts, or situations in some knowledge domain, and links represent the relations, associations, or dependencies between them. Semantic networks have been used as cognitive models of human memory and representational schemes for artificial intelligence systems. These symbolic networks are essentially universal or generic in nature. Being applicable or suitable to an almost infinite number of knowledge domains or subject matters, it makes sense in terms of minimizing effort and cost to develop computer-based testing systems that incorporate semantic networks. However, an important question remains to be answered: How effective are these systems when compared to more customary measurement methods? Differences between computer-based assessment employing semantic networks and paper-based traditional testing techniques may or may not impact upon the reliability and validity of measurement. The primary purpose of this reported research was to shed some light on this salient issue by evaluating empirically the relative reliability and validity of a computer-based and a paper-based procedure for assessing semantic knowledge.

METHOD

Subjects

The subjects were 75 male F-14 pilots, radar intercept officers (RIOs), and students, as well as E-2C pilots and naval flight officers (NFOs) from training and operational squadrons at Naval Air Station (NAS) Miramar. All had volunteered to participate in this research.

Subject Matter

A database was developed that consisted of five categories of facts about front-line Soviet platforms: (a) weapons systems, (b) radar and electronic countermeasure (ECM) systems, (c) surface and subsurface platforms, (d) airborne platforms, and (e) counterjamming procedures. It was used to train and test the subjects concerning important threat parameters associated with Russian platforms (e.g., aircraft range and speed; payload of antiship missiles; typical launch altitude; missile range, flight profile, velocity, and warheads; other weapon, radar, ECM/ECCM [electronic counter-countermeasure] systems; surveillance capabilities).

The database was structured as a semantic network (Barr & Feigenbaum, 1981; Johnson-Laird, 1983) in order to represent the associative knowledge inherent to it for computer systems. That is, objects and their corresponding properties, attributes, or characteristics were represented as node-link structures. The links between nodes represent the associations or relationships among objects or among objects and their attributes. For example, the object "aircraft type" and the attribute "ECM suite" can be linked so that the system can represent a particular aircraft

type that has a certain ECM suite. By defining initially all objects and attributes in the database, a hierarchy or tree structure can be specified for all objects, attributes, and their relationships. Once a database was structured as a semantic network, it became possible for independent software modules to interact with, operate upon, or manipulate the database. For example, interpretative programs could ask questions about the database, since its intrinsic structure was represented. This latter capability was capitalized upon in this research.

Computer-Based Assessment

A computer-based game or test, FlashCards (Liggett & Federico, 1986), was adopted and adapted to quiz students and instructors as well as crew members of other operational squadrons about the threat-parameter database. This computer-based quiz is totally autonomous or independent of the database and will run on any database structured as a semantic network. It randomly selects objects or chooses characteristics from the database, and generates questions about threat platforms or their salient attributes. Unlike some computer-based tests, alterative forms did not have to be specifically or previously programmed as such.

FlashCards is analogous to using real flash cards. That is, a question is presented to individual students who are expected to answer it. Questions can have multiple answers as with "What Soviet bombers carry the XYZ-123 missile?" After individual students are presented with the question, they are allowed as many tries as they would like to type in the answer. If the students cannot answer the question, they can continue with the quiz. At this point, they are provided feedback in terms of the correct answer or answers. At any point in the answering process, they can continue to the next question. For each answer, the students must key in a response which reflects their degree of confidence in their answer. Also, for each answer the student's response latency is recorded and displayed.

FlashCards quizzed the students on all top-level, or general, categories of the semantic network that it was using as the database. The score for each question was computed as the number of correct answers entered divided by the total number of answers entered. For the purposes of this research, a FlashCards test consisted of 25 completion or fill-in-the-blank domain-referenced items or questions. These were considered as two groups of 12 odd and even items each (dropping the last question) for computing split-half reliability estimates. The average score for odd (even) items was calculated as the total score of odd (even) items divided by the number of odd (even) questions attempted. The total computer-based test score was calculated as the average of the odd and even halves.

Paper-Based Assessment

Two alterative forms of a paper-based test were designed and developed to assess knowledge of the same threat-parameter database mentioned above, and to mimic as much as possible the format used by FlashCards. Both of these consisted of 25 completion or fill-in-the-blank domain-referenced items or questions. As with the computer-based test, more than one answer may be required per item or question. Beneath each question was a confidence scale that resembled the one used in FlashCards where the test-takers were required to indicate the level of confidence in their response(s). Scoring items for this paper-based test was similar to scoring the computer-based test: For each question, the number of correct answers given was divided by the total number of answers completed for that question. Also, scoring odd (even) halves of the test for computing internal consistency was simi-

lar to that for FlashCards. The score for the total paper-based test was calculated like the total score for the computer-based test.

Procedure

Subjects acquired threat-parameter knowledge using dual media: (a) a traditional text organized according to the database's major topics and (b) a computer-based system consisting of the quizzes FlashCards and Jeopardy. Mode of assessment, computer-based or paper-based, was manipulated as a within-subjects variable (Kirk, 1968). Subjects were administered the computer-based and paper-based tests in counterbalanced order. The two forms of the paper-based tests were alternated in their administration to subjects. After subjects received either the computer-based or paper-based test, they were immediately administered the other. It was assumed that a subject's state of threat-parameter knowledge was the same during the administration of both tests. Subjects took approximately 10–15 min to complete the paper-based test, and 20–25 min to complete the computer-based test. The longer time to complete the latter test was largely attributed to lack of typing or keyboard proficiency on the part of some of the subjects. The manner in which the subject matter was presented during assessment within the computer-based and paper-based media was essentially the same, due to similar symbol systems and presentation formats being employed.

Reliabilities for both modes of testing were estimated by deriving internal consistency indices using an odd–even item split. These reliability estimates were adjusted by employing the Spearman–Brown Prophecy Formula (Thorndike, 1982). Reliability estimates were calculated for test score, average degree of confidence, and average response latency for the computer-based test, but only for test score and average degree of confidence for the paper-based test. Response latency was not measured for the paper-based test. Equivalences between the two modes of assessment were estimated by Pearson product–moment correlations for total test score and average degree of confidence. These correlations were considered indices of the extent to which the two types of testing were measuring the same semantic knowledge and amount of assurance in answers.

In order to derive discriminant validity estimates, research subjects were placed into four groups: those above or below F-14 or E-2C mean flight hours. One stepwise multiple discriminant analysis, using Wilks' criterion for including and rejecting variables and their associated statistics, was computed to ascertain how well computer-based and paper-based measures distinguished among the defined groups that were expected to differ in the extent of their knowledge of the threat-parameter database. It was thought that mean flight hours reflect operational experience. Those individuals with more operational experience were expected to perform better on tests of threat-parameter knowledge than those with less experience. Also, F-14 crew members were expected to have more knowledge of specific threat parameters than E-2C crew members since the former must be intimately more familiar with these attributes in order to make successful intercepts than the latter.

RESULTS

Reliability and Equivalence Estimates

Split-half reliability and equivalence estimates of computer-based and paper-based measures from the pooled within-groups correlation matrices for the different

groupings are tabulated in Table 1. It can be seen that the adjusted reliability estimates of the computer-based and paper-based measures are moderate to high, ranging from .74 to .97. None of the differences in corresponding reliabilities for computer-based and paper-based measures, test score and average degree of confidence, was found to be statistically significant ($p > .01$) using a test described by Edwards (1964). This suggests that the computer-based and paper-based measures were not significantly different in reliability or internal consistency.

Considering the computer-based measures, it was ascertained that the reliability estimate for average degree of confidence was significantly ($p < .01$) higher than the reliability estimates for average response latency and test score. Also, the reliability estimate for response latency was significantly higher than the one computed for test score. Focusing on the paper-based measures, it was found that the reliability estimate for average degree of confidence was significantly ($p < .01$) higher than the reliability estimate for test score. These results implied that these measures can be ranked in order of their internal consistencies from highest to lowest as follows: average degree of confidence, average response latency, and test score.

Equivalence estimates for test score and average degree of confidence measures, respectively, were .76 and .82. These suggest that the computer-based and paper-based measures had anywhere from approximately 58 to 67% variance in common, implying that these different modes of assessment were somewhat or partially equivalent. The equivalences for test score and average degree of confidence measures were not significantly ($p > .01$) different.

Discriminant Validity Estimates

The discriminant analysis computed to determine how well computer-based and paper-based measures differentiated groups defined by above or below F-14 or E-2C mean flight hours yielded one significant discriminant function. According to the multiple discriminant analysis model (Cooley & Lohnes, 1962; Tatsuoka, 1971; Van de Geer, 1971), the maximum number of derived discriminant functions is either one less than the number of groups or equal to the number of discriminating variables, whichever is smaller. Since there were four groups to be discriminated, this analysis yielded three discriminant functions, but only one of them was significant. Consequently, solely this significant discriminant function and its associated statistics are presented.

The statistics associated with the significant function, standardized discriminant-function coefficients, pooled within-groups correlations between the function and computer-based and paper-based measures, and group centroids for above or below

Table 1. Split-Half Reliability and Equivalence Estimates of Computer-Based and Paper-Based Measures for Semantic Knowledge

	Reliability		
Measure	Computer-Based	Paper-Based	Equivalence
Score	.74	.76	.76
Confidence	.96	.97	.82
Latency	.88	—	—

Note Split-half reliability estimates were adjusted by employing the Spearman–Brown Prophecy Formula

F-14 or E-2C mean flight hours are presented in Table 2. It can be seen that the single significant discriminant function accounted for approximately 82% of the variance among the four groups. The discriminant-function coefficients that consider the interrelationships or interdependencies among the multivariate measures revealed the relative contribution or comparative importance of these variables in defining this derived dimension to be the paper-based total score (PTS), the computer-based total score (CTS), and the computer-based total average degree of confidence (CTC), respectively. The computer-based total average latency (CTL) and the paper-based total average degree of confidence (PTC) were considered unimportant in specifying this discriminant function since the absolute values of their coefficients were each below .4. The within-groups correlations that are computed for each individual measure partialling out the interrelationships of all the other variables indicated that the major contributors to the significant discriminant function were CTC, CTS, and CTL, respectively, all computer-based measures. The group centroids showed how the performance of the F-14 crew members clustered together along one end of the derived dimension, while the performance of the E-2C crew members clustered together along the other end of the continuum. The means and standard deviations for groups above or below F-14 or E-2C mean flight hours, univariate F ratios, and levels of significance for computer-based and paper-based measures are tabulated in Table 3. Considering the measures as univariate variables — that is, independent of their multivariate relationships with one another — these statistics revealed that the three computer-based measures CTC, CTS, and CTL, respectively, significantly differentiated the four groups, not the paper-based measures, PTS and PTC. Applying Duncan's multiple range test (Kirk, 1968) on the group means for the important individual measures indicated that F-14 crews significantly ($p < .05$) outperformed E-2C crews on CTS, CTC, and CTL. The multivariate and subsequent univariate results established the discriminant validity of computer-based measures to be superior to that of paper-based measures.

Table 2. Statistics Associated With Significant Discriminant Function, Standardized Discriminant-Function Coefficients, Pooled Within-Groups Correlations Between the Discriminant Function and Computer-Based and Paper-Based Measures, and Group Centroids for Above or Below F-14 or E-2C Mean Flight Hours

Discriminant Function						
Eigenvalue	Percent Variance	Canonical Correlation	Wilks Lambda	Chi-Square	*df*	*p*
.44	82.43	.55	.64	31.38	15	.008
Measure	Discriminant Coefficient	Within-Group Correlation		Group	Centroid	
CTS	.91	.51		Above F-14 Mean Hours	.10	
CTC	.84	.57		Below F-14 Mean Hours	.39	
CTL	−.24	−.45		Above E-2C Mean Hours	−1.35	
PTS	−1.19	−.00		Below E-2C Mean Hours	−1.50	
PTC	−.17	.36				

Note: CTS = Computer-based total test score. CTC = average degree of confidence. CTL = average response latency. PTS = paper-based total test score. PTC = average degree of confidence. CTL was measured in seconds.

Table 3. Means and Standard Deviations for Groups Above or Below F-14 or E-2C Mean Flight Hours, Univariate *F* Ratios, and Levels of Significance for Computer-Based and Paper-Based Measures

Measure		Group					
		Above F-14 Flight Hours (n = 26)	Below F-14 Flight Hours (n = 37)	Above E-2C Flight Hours (n = 5)	Below E-2C Flight Hours (n = 7)	F	p
CTS							
	M	60.58	59.62	44.60	43.14	2.94	.039
	SD	15.75	18.77	15.68	17.37		
CTC							
	M	75.58	80.84	48.60	64.57	4.11	.010
	SD	21.57	19.80	21.23	26.48		
CTL							
	M	8.42	7.81	9.49	11.06	2.28	.087
	SD	3.31	2.77	4.10	3.94		
PTS							
	M	51.65	49.73	45.80	52.86	.19	.900
	SD	18.26	20.38	11.86	13.91		
PTC							
	M	72.23	76.70	53.0[illegible]	69.71	2.14	.103
	SD	23.02	18.10	[illegible]	20.94		

DISCUSSION

This study established that (a) computer-based and paper-based measures, test score and average degree of confidence, are not significantly different in reliability or internal consistency; (b) for computer-based and paper-based measures, average degree of confidence has a higher reliability than test score; (c) the equivalence estimates for computer-based and paper-based measures (test score and average degree of confidence) were not significantly different; and (d) the discriminant validity of the computer-based measures was superior to paper-based measures.

The finding that computer-based and paper-based measures, test score and average degree of confidence, were not significantly different in reliability or internal consistency partially agrees with the corresponding result established in a study by Federico (1991). In that research, computer-based and paper-based measures of test scores for recognition of aircraft silhouettes were found to be equally reliable; however, the computer-based measure of average degree of confidence was found to be less reliable than its paper-based counterpart. The present study suggested that equivalence estimates for computer-based and paper-based measures, test score and average degree of confidence, were dissimilar in magnitude. This finding is similar to that established in the Federico (1991) study where computer-based and paper-based measures of test score were less equivalent than these measures of average degree of confidence. Lastly, some of the results of the present research demonstrated that the discriminative validity of the computer-based measures was superior to paper-based measures. This finding is in partial agreement with that found in the Federico (1991) research where this was also established with respect to some statistical criteria. However, according to other criteria the discriminative validities of computer-based and paper-based measures were about the same.

Computer-based and paper-based media vary in the nature of the reciprocal interaction and information feedback they provide to individuals during learning or

testing. Isolating on assessment, usually, computer-based media provide more immediate interaction and information feedback to the test-taker than paper-based media. Typically, the computer system presents a question that an individual attempts to answer, and the quality of the response is immediately displayed. Or, in computerized adaptive or tailored testing the system presents a test item and, based upon the correctness of the individual's response, then provides either a more or a less difficult follow-on item. That is, the computer-based system is interactive to the degree that it is designed to tailor, or adapt, the level of difficulty of the administered items as a function of test-takers' responses. In these contexts, the direct interaction or continuous transaction between the test-taker and the system is intrinsic to the establishment of the feedback loop, which was the case with the computer-based assessment system used in this reported research.

Three distinct functions have been attributed to feedback, namely: (a) reinforcement, (b) information, and (c) motivation (Bilodeau, 1966). Each of these three attributes is more apparent in computer-based than in paper-based assessment. Computer-based testing or gaming systems can be designed to reinforce directly correct responses by awarding a number of points to the test-taker or player. It is difficult, though not impossible, for paper-based testing to match the promptness of the reinforcement provided by computer-based testing. Usually, the immediacy of the information provided by a computer-based system concerning the correctness or incorrectness of a response to a test item exceeds that provided by a paper-based system. Partly due to the almost simultaneous administration of reinforcement and display of information as a direct consequence of responding, the level of motivation typically elicited by a computer-based system should surpass that aroused by a paper-based system. Also, some computer-based quizzes are essentially game-like in nature, like the ones employed in this reported research and in Federico's (1991) study. The incentive provided by assessment systems such as these approaches that of video games where players attempt to outperform one another by maximizing their individual payoffs. The desire to establish a personal best, to surpass the others, or to be in the top ten is instilled by using some well-designed computer-based testing systems and/or because of the mere fact that individual performance can be visible to others when interacting with this video game-like technology, thus eliciting socially motivated competitive behavior. This desire testifies to the typically higher level of engagement experienced by people when employing computer-based than paper-based assessment systems.

Computer-based testing or gaming systems usually have as an integral component video display terminals that are similar to television screens. Consequently, people possibly perceive, expect, or anticipate a priori that assessment systems such as these may be more engaging, engrossing, or entertaining than paper-based tests or games, regardless of the subject-matter domain. That is, personal perceptions or expectations concerning the measurement system as well as the assessment situation predispose how tests are taken by individuals. Within the current zeitgeist, it seems reasonable to expect that people generally have more positive attitudes toward computer-based than paper-based media, partly because of the perceived higher entertainment potential of the former. The associative, affective, and active tendencies attributed to these stronger positive attitudes may culminate in people perceiving computer-based media as more absorbing and attracting than paper-based media. That is, in this high technology era individuals seem more interested in, or inclined toward, attending to or heeding computer-based rather than paper-based media.

Extrapolating from this implicit framework, or engagement theory, it was expected that the computer-based test used in this study would have higher reliabil-

ities and validities than the paper-based test, regardless of the measurement criteria employed, because the former should have provided more immediate interaction and information feedback, instilled a higher level of motivation, and engaged individuals to a greater extent than the latter. More interaction, feedback, motivation, or engagement evoked by the computer-based test should have encouraged or exhorted individuals to exert or energize their performances during measurement maximally and continuously. That is, subjects were expected to amplify their respective performances, because of the greater expected engagement elicited in them, when interacting with a computer-based rather than a paper-based test. Subjects should have consistently or continuously sustained their maximum efforts throughout the entire computer-based test, culminating in possibly less response variability, and consequently more reliability, than the paper-based test. Higher reliability, in turn, should have resulted in higher discriminative validity for computer-based than for paper-based measurement. This was not entirely and empirically established by this reported research. Contrary to what was implicitly expected, this study demonstrated that the reliabilities of computer-based and paper-based tests are not significantly different. Compatible with the presumed framework, however, this investigation found that computer-based measures had validity superior to paper-based measures.

Hofer and Green (1985) were concerned that computer-based assessment would introduce irrelevant or extraneous factors that would likely degrade test performance. These computer-correlated factors may alter the nature of the task to such a degree that it would be difficult for a computer-based test and its paper-based counterpart to measure the same construct or content. This could impact upon reliability, validity, and normative data, as well as other assessment attributes. Several plausible reasons, they stated, may contribute to different performances on these distinct kinds of testing: (a) state anxiety instigated when confronted by computer-based testing, (b) lack of computer familiarity on the part of the test-taker, and (c) changes in response format required by the two modes of assessment. These different dimensions could result in tests that are nonequivalent; however, in this reported research these diverse factors had no apparent impact.

On the other hand, there are a number of known differences between computer-based and paper-based assessment that may affect equivalence and validity (Green, 1986):

1. Passive omitting of items is usually not permitted on computer-based tests. An individual must respond, unlike with most paper-based tests.
2. Computerized tests typically do not permit backtracking. The test-taker cannot easily review items, alter responses, or delay answering questions.
3. The capacity of the computer screen can have an impact on what usually are long test items (e.g., paragraph comprehension). These may be shortened to accommodate the computer display, thus partially changing the nature of the task.
4. The quality of computer graphics may affect the comprehension and degree of difficulty of the item.
5. Pressing a key or using a mouse is probably easier than marking an answer sheet. This may impact upon the validity of speeded tests.
6. The computer typically displays items individually; traditional time limits are no longer necessary.

Assuming that these abstract distinctions may affect the equivalence and validity of computer-based and paper-based assessment, the omission of items and back-

tracking on paper-based tests in this research was not permitted in order to simulate computer-based tests. Computer screen capacity was of no consequence in this study since none of the test items was long. Graphics were not employed in the paper-based test or its computer-based counterpart and consequently played no part in item comprehension or difficulty. In this study neither the computer-based nor paper-based measurement employed true speeded tests. Also, to mimic the individual display of items on the computer-based tests, the subjects were closely monitored as they took the paper-based test, and were reminded to expedite their responses without retracing.

When evaluating or comparing different media for instruction and assessment, one must keep in mind that the newer medium may simply be seen as more interesting, engaging, and challenging by the students. This novelty effect seems to disappear as rapidly as it appears. However, in research studies conducted over a relatively short time span, for example, a few days or months at the most, this effect may still linger and affect the evaluation by its enhancement of the impact of the more novel medium (Colvin & Clark, 1984), which could have occurred in this reported research. When matching media to distinct subject matters, course contents, or core concepts, some research evidence (Jamison, Suppes, & Welles, 1974) indicates that, other than in obvious cases, just about any medium will be effective for different content. Extrapolating this notion to the measurement domain, the validity results of this study seemed to suggest, contrary to the above, that different media may be differentially effective testing of the same subject matter.

Acknowledgment — The assistance of Nina Liggett Bacas and Janice Singer is appreciated and acknowledged.

REFERENCES

Barr, A., & Feigenbaum, E. F. (Eds.). (1981). *The handbook of artificial intelligence (Vol. 1)*. Stanford, CA: HeurisTech.

Bilodeau, I. McD. (1966). Information feedback. In E. A. Bilodeau (Ed.), *Acquisition of skill*. New York: Academic Press.

Colvin, C., & Clark, R. E. (1984, July). Instructional media vs. instructional methods. *Performance and Instruction Journal*, pp. 1–3.

Cooley, W. W., & Lohnes, P. R. (1962). *Multivariate procedures for the behavioral sciences*. New York: Wiley.

Cory, C. H. (1977). Relative utility of computerized versus paper-and-pencil tests for predicting job performance. *Applied Psychological Measurement*, **1**, 551–564.

Divgi, D. R. (1988, October). *Two consequences of improving a test battery* (CRM 88-171). Alexandria, VA: Center for Naval Analyses.

Edwards, A. L. (1964). *Experimental design in psychological research*. New York: Holt, Rinehart, and Winston.

Elwood, D. L., & Griffin, R. H. (1972). Individual intelligence testing without the examiner: Reliability of an automated method. *Journal of Consulting and Clinical Psychology*, **38**, 9–14.

English, R. A., Reckase, M. D., & Patience, W. M. (1977). Applications of tailored testing to achievement measurement. *Behavior Research Methods & Instrumentation*, **9**, 158–161.

Evan, W. M., & Miller, J. R. (1969). Differential effects of response bias of computer versus conventional administration of a social science questionnaire. *Behavioral Science*, **14**, 216–227.

Federico, P.-A. (1980). Adaptive instruction: Trends and issues. In R. E. Snow, P.-A. Federico, & W. E. Montague (Eds.), *Aptitude, learning, and instruction: Vol. 1. Cognitive process analyses of aptitude*. Hillsdale, NJ: Erlbaum.

Federico, P.-A. (1991). Measuring recognition performance using computer-based and paper-based methods. *Behavior Research Methods, Instruments, & Computers*, **23**, 341–347.

Green, B. F. (1983a). Adaptive testing by computer. *Measurement, Technology, and Individuality in Education*, **17**, 5–12.

Green, B. F. (1983b). The promise of tailored tests. *Principles of modern psychological measurement: A festschrift in honor of Frederic Lord*. Hillsdale, NJ: Erlbaum.

Green, B. F. (1986). *Construct validity of computer-based tests*. Paper presented at the Test Validity Conference, Educational Testing Service, Princeton, NJ.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, **21**, 347–360.

Hansen, D. H., & O'Neil, H. F. (1970). Empirical investigations versus anecdotal observations concerning anxiety and computer-assisted instruction. *Journal of School Psychology*, **8**, 315–316.

Hedl, J. J., O'Neil, H. F., & Hansen, D. H. (1973). Affective reactions toward computer-based intelligence testing. *Journal of Consulting and Clinical Psychology*, **40**, 217–222.

Hitti, F. J., Riffer, R. L., & Stuckless, E. R. (1971, July). *Computer-managed testing: A feasibility study with deaf students*. Rochester, NY: National Technical Institute for the Deaf.

Hofer, P. J., & Green, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology*, **53**, 826–838.

Hoffman, K. I., & Lundberg, G. D. (1976). A comparison of computer-monitored group tests with paper-and-pencil tests. *Educational and Psychological Measurement*, **36**, 791–809.

Jamison, D., Suppes, P., & Welles, S. (1974). The effectiveness of alternative media: A survey. *Annual Review of Educational Research*, **44**, 1–68.

Johnson, D. F., & Mihal, W. L. (1973). Performance of blacks and whites in computerized versus manual testing environments. *American Psychologist*, **28**, 694–699.

Johnson, D. F., & White, C. B. (1980). Effects of training on computerized test performance in the elderly. *Journal of Applied Psychology*, **65**, 357–358.

Johnson, J. H., & Johnson, K. N. (1981). Psychological considerations related to the development of computerized testing stations. *Behavior Research Methods & Instrumentation*, **13**, 421–424.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Kantor, J. (1988). *The effects of anonymity, item sensitivity, trust, and method of administration on response bias on the job description index*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Katz, L., & Dalby, J. T. (1981). Computer-assisted and traditional psychological assessment of elementary-school-age children. *Contemporary Educational Psychology*, **6**, 314–322.

Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Brooks/Cole.

Koson, D., Kitchen, C., Kochen, M., & Stodolosky, D. (1970). Psychological testing by computer. Effect on response bias. *Educational and Psychological Measurement*, **30**, 808–810.

Lee, J. A., Moreno, K. E., & Sympson, J. B. (1984, April). *The effects of mode of test administration on test performance*. Paper presented at the annual meeting of the Eastern Psychological Association, Baltimore.

Liggett, N. L., & Federico, P.-A. (1986). *Computer-based system for assessing semantic knowledge: Enhancements* (NPRDC TN 87-4). San Diego, CA: Navy Personnel Research and Development Center.

Lucas, R. W., Mullin, P. J., Luna, C. D., & McInroy, D. C. (1977). Psychiatrists and a computer as interrogators of patients with alcohol related illnesses: A comparison. *British Journal of Psychiatry*, **131**, 160–167.

Lukin, M. E., Dowd, E. T., Plake, B. S., & Kraft, R. G. (1985). Comparing computerized versus traditional psychological assessment. *Computers in Human Behavior*, **1**, 49–58.

Lushene, R. E., O'Neil, H. F., & Dunn, T. (1974). Equivalent validity of a completely computerized MMPI. *Journal of Personality Assessment*, **34**, 353–361.

McBride, J. R. (1980). Adaptive verbal ability testing in a military setting. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology.

Millman, J. (1984, Summer). Using microcomputers to administer tests: An alternate point of view. *Educational Measurement: Issues and Practices*, pp. 20–21.

Rock, D. L., & Nolen, P. A. (1982). Comparison of the standard and computerized versions of the raven coloured progressive matrices test. *Perceptual and Motor Skills*, **54**, 40–42.

Sachar, J. D., & Fletcher, J. D. (1978). Administering paper-and-pencil tests by computer, or the medium is not always the message. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology

Serwer, B. L., & Stolurow, L. M. (1970). Computer-assisted learning in language arts. *Elementary English*, **47**, 641–650.

Skinner, H. A., & Allen, B. A. (1983). Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug, and tobacco use. *Journal of Consulting and Clinical Psychology*, **51**, 267–275.

Space, L. G. (1981). The computer as psychometrician. *Behavior Research Methods & Instrumentation*, **13**, 595–606.

Sympson, J. B., Weiss, D. J., & Ree, M. (1982). *Predictive validity of conventional and adaptive tests in an air force training environment* (AFHRL-TR-81-40). Brooks AFB, TX: Air Force Human Resources Laboratory.

Tatsuoka, M. M. (1971). *Multivariate analysis*. New York: Wiley.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.

Van de Geer, J. P. (1971). *Introduction to multivariate analysis for the social sciences*. San Francisco: W. H. Freeman.

Ward, W. C. (1984, Summer). Using microcomputers to administer tests. *Educational Measurement: Issues and Practices*, pp. 16–20.

Wildgrube, W. (1982, July). *Computerized testing in the German federal armed forces — empirical approaches*. Paper presented at the 1982 Computerized Adaptive Testing Conference, Spring Hill, MN.